OPTIMAL LINEAR ESTIMATION OF BOUNDS OF RANDOM VARIABLES

BY

PETER COOKE

TECHNICAL REPORT NO. 37
SEPTEMBER 24, 1979

PREPARED UNDER GRANT
DAAG29-77-G-0031
FOR THE U.S. ARMY RESEARCH OFFICE

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Optimal Linear Estimation of Bounds of Random Variables

By

Peter Cooke

TECHNICAL REPORT NO. 37

September 24, 1979

Prepared under Grant DAAG29-77-G-0031
For the U.S. Army Research Office

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA

Partially supported under Office of Naval Research Contract N00014-76-C-0475 (NR-042-267) and issued as Technical Report No. 276.

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Optimal Linear Estimation of Bounds of Random Variables

By

Peter Cooke

1. Introduction.

Suppose $X_1, X_2, \ldots, X_n$ are independent random variables, each with density $f(x)$ and distribution function $F(x)$, where $F(x) \in (0,1)$ only for $x \in (\varphi, \theta)$. Let $Y_1 \leq Y_2 \leq \cdots \leq Y_n$ denote the order statistics based on $X_1, X_2, \ldots, X_n$. The parameter θ is known to be finite and is the parameter of interest. The large sample inference for θ which follows applies whether or not φ is known, though when $\varphi = -\infty$ we will need to assume that the X's have finite second moment since our estimators are linear functions of the order statistics and we will compare estimators through their mean squared errors.

When only the r largest observations are used to estimate θ, a linear estimator is of the form

$$\hat{\theta}_{n,r} = \sum_{i=1}^{r} a_i Y_{n-i+1} \ . \tag{1}$$

In section 2 we will show, for fixed $r \geq 2$, how the coefficients $a_1, a_2, \ldots, a_r$ can be determined so as to yield the estimator of the form (1) with asymptotically smallest mean squared error.

2. Determination of the Coefficients.

It is clear that since we are discussing large sample theory and the parameter of interest is the upper endpoint of the distribution and also, since we are basing our inference on the largest few observations, from a practical point of view we don't need to know the form of f, but we need to characterize the shape of its upper tail. Thus, as in Cooke (1979) we will consider the case in which

$$F^n(y) \sim \exp\{-(\frac{\theta-y}{\theta-u_n})^{1/\nu}\} \text{ as } n \to \infty , \tag{2}$$

for which Gnedenko's (1943) necessary and sufficient condition is that for $c > 0$,

$$\lim_{y \to 0-} \frac{1-F(cy+\theta)}{1-F(y+\theta)} = c^{1/\nu} , \quad \text{where} \quad u_n = F^{-1}(1-\frac{1}{n}) .$$

The value $\nu = 1$ corresponds to densities $f(x)$ which are truncated at θ; that is, $0 < f(\theta) < \infty$. In general, $\nu = 1/(k+1)$ for a density which is zero or infinite at θ and whose first finite, nonzero left derivative at θ is its k^{th} left derivative.

It is proved in Cooke (1979) that, when Gnedenko's condition holds and $n \to \infty$, for $i \geq 1$ and i small

$$E(Y_{n-i+1}) \sim \theta-(\theta-u_n) \frac{\Gamma(\nu+i)}{\Gamma(i)} \tag{3}$$

and, for $i \geq j \geq 1$,

$$\text{(4)} \qquad \operatorname{Cov}(Y_{n-i+1}, Y_{n-j+1}) \sim \frac{\Gamma(\nu+j)}{\Gamma(j)} \left\{\frac{\Gamma(2\nu+i)}{\Gamma(\nu+i)} - \frac{\Gamma(\nu+i)}{\Gamma(i)}\right\} ,$$

where $\Gamma(\alpha)$ is the familiar gamma function defined by

$$\Gamma(\alpha) = \int_0^\infty e^{-x} x^{\alpha-1} dx \quad \text{for} \quad \alpha > 0 .$$

It follows from (3) that as $n \to \infty$

$$\text{(5)} \qquad E(\hat{\theta}_{n,r} - \theta) \sim \theta\left(\sum_{i=1}^{r} a_i - 1\right) - (\theta - u_n) \sum_{i=1}^{r} a_i \frac{\Gamma(\nu+1)}{\Gamma(i)} .$$

When $\sum\limits_{i=1}^{r} a_i = 1$, which we require for $\hat{\theta}_{n,r}$ to be a consistent estimator of θ, using (4) and (5) we find, when $n \to \infty$,

$$\text{(6)} \qquad (\theta - u_n)^{-2} E(\hat{\theta}_{n,r} - \theta)^2 \sim \sum_{i=1}^{r} \sum_{j=1}^{r} a_i a_j \frac{\Gamma(2\nu+i)\Gamma(\nu+j)}{\Gamma(\nu+i)\Gamma(j)} .$$

The quadratic form on the right in (6) can be written as $\underset{\sim}{a}' \Lambda \underset{\sim}{a}$, where $\underset{\sim}{a}$ is a column vector with elements $a_1, a_2, \ldots, a_r$ and Λ is a symmetric $r \times r$ matrix with $(i,j)^{th}$ element

$$\lambda_{ij} = \frac{\Gamma(2\nu+i)\Gamma(\nu+j)}{\Gamma(\nu+i)\Gamma(j)} , \quad j \leq i .$$

If we let $\underset{\sim}{1}$ denote the $r \times 1$ vector with each element equal to 1, our problem reduces to finding the vector $\underset{\sim}{a}$ which minimizes $\underset{\sim}{a}' \Lambda \underset{\sim}{a}$ subject to $\underset{\sim}{a}'\underset{\sim}{1} = 1$. The minimization is achieved by the vector $\underset{\sim}{a} = (\underset{\sim}{1}' \Lambda^{-1} \underset{\sim}{1})^{-1} \Lambda^{-1} \underset{\sim}{1}$ and the minimum value of $\underset{\sim}{a}' \Lambda \underset{\sim}{a}$ is $(\underset{\sim}{1}' \Lambda^{-1} \underset{\sim}{1})^{-1}$.

3. Computations.

In the tables to follow we have, correct to three decimal places, values of the coefficients of the r largest order statistics for the minimum mean squared error estimator of θ, which henceforth we denote by $\hat{\theta}_{n,r}$. Also tabulated are some values of $\gamma_r(\nu) = \lim_{n \to \infty} (\theta-u_n)^{-2}E(\hat{\theta}_{n,r}-\theta)^2$.

The truncation case $\nu = 1$ is probably the most important case from a practical point of view, but is singled out here in view of the special nature of the minimizing coefficients $a_1,a_2,\ldots,a_r$. Gnedenko's condition suggests that for y close to θ, $1-F(y) \propto (\theta-y)^{1/\nu}$ and hence, when $\nu = 1$, that $F(y)$ is linear in y for y near θ. This corresponds to a Uniform distribution. If indeed $Y_{n-r+1},Y_{n-r+2},\ldots,Y_n$ are the r largest order statistics from a Uniform distribution with upper endpoint θ and $Y_1,Y_2,\ldots,Y_{n-r}$ are ignored, then Y_{n-r+1} and Y_n are jointly sufficient for θ, in which case it follows that the minimum mean squared error estimator of θ will be a linear function of Y_{n-r+1} and Y_n alone. Thus $a_2 = a_3 = \cdots = a_{r-1} = 0$ when $\nu = 1$. The increasing dependence on $Y_{n-r+2},Y_{n-r+3},\ldots,Y_{n-1}$ with decreasing ν or, equivalently, increasing power of $(\theta-y)$, is apparent from the tables to follow.

Using (6) with $\nu = 1$ and $a_2 = a_3 = \cdots = a_{r-1} = 0$ we easily find that the minimizing coefficients are $a_1 = 1+r^{-1}$, $a_r = -r^{-1}$ and that the minimum value of $\gamma_r(1)$ is $1+r^{-1}$. It follows that $\gamma_r(1)$ cannot be smaller than 1 for any $r \geq 1$ and that a nearly optimal estimator is obtained with a fairly small value of r.

Table 1

Minimizing Coefficients and Asymptotic Mean Squared Error of the Optimal Estimator.

$\nu = 1/2$

r	a_1	a_2	a_3	a_4	a_5	a_6	a_7	$\gamma_r(1/2)$
2	2	-1						.667
3	1.636	.273	-.909					.545
4	1.440	.240	.160	-.840				.480
5	1.314	.219	.146	.109	-.788			.438
6	1.224	.204	.136	.102	.082	-.748		.408
7	1.157	.193	.129	.096	.077	.064	-.716	.386

$\nu = 1/3$

r	a_1	a_2	a_3	a_4	a_5	a_6	a_7	$\gamma_r(1/3)$
2	2.5	-1.5						.564
3	1.951	.585	-1.537					.440
4	1.654	.496	.372	-1.523				.373
5	1.463	.439	.329	.269	-1.501			.330
6	1.328	.398	.299	.244	.210	-1.479		.300
7	1.226	.368	.276	.226	.193	.171	-1.459	.277

$\nu = 1/4$

r	a_1	a_2	a_3	a_4	a_5	a_6	a_7	$\gamma_r(1/4)$
2	3	-2						.532
3	2.273	.909	-2.182					.403
4	1.882	.753	.602	-2.237				.334
5	1.632	.653	.522	.448	-2.255			.289
6	1.456	.583	.466	.399	.355	-2.260		.258
7	1.325	.530	.424	.363	.323	.294	-2.259	.235

Table 1 (Continued)

Minimizing Coefficients and Asymptotic Mean Squared Error of the Optimal Estimator.

$\nu = 1/5$

r	a_1	a_2	a_3	a_4	a_5	a_6	a_7	$\gamma_r(1/5)$
2	3.5	-2.5						.518
3	2.598	1.237	-2.835					.384
4	2.117	1.008	.840	-2.964				.313
5	1.811	.863	.719	.634	-3.027			.268
6	1.598	.761	.634	.560	.509	-3.062		.236
7	1.439	.685	.571	.504	.458	.424	-3.082	.213

Although the minimizing coefficients are not given above for $r = 20$, except when $\nu = 1$, the following table gives values of $\eta_{20}(\nu)$, where

$$\eta_r(\nu) = \lim_{n \to \infty} \frac{E(\hat{\theta}_{n,r}-\theta)^2}{E(\bar{\theta}_n-\theta)^2}$$

is the asymptotic efficiency of $\bar{\theta}_n$ relative to $\hat{\theta}_{n,r}$ and, as discussed in Cooke (1979), $\bar{\theta}_n$ is the estimator of the form

$$Y_n + c(\nu)\{Y_n-(1-e^{-1}) \sum_{i=0}^{n-1} e^{-i} Y_{n-i}\}$$

with asymptotically smallest mean squared error and is the best estimator derived until now. Also tabulated are values of $\delta_{20}(\nu)$, where

$$\delta_r(\nu) = \lim_{n \to \infty} \frac{E(\hat{\theta}_{n,r}-\theta)^2}{E(Y_n-\theta)^2} ,$$

to illustrate the considerable progress which has been made in finding better estimators of θ than Y_n since Robson and Whitlock's (1964) attempt in the truncation case.

Table 2

Efficiencies Relative to the Optimal Estimator Based on the 20 Largest Observations and Improvement Over Y_n.

ν	1	1/2	1/3	1/4	1/5
$\eta_{20}(\nu)$	.798	.494	.357	.257	.252
$\delta_{20}(\nu)$	.525	.278	.185	.143	.120

4. Estimation of φ.

When φ is known to be finite and is the parameter of interest, for given $r \geq 1$ we seek the estimator of the form $\hat{\varphi}_{n,r} = \sum_{i=1}^{r} a_i Y_i$ with asymptotically smallest mean squared error.

If the lower tail of f is characterized by the constant ν in the way the upper tail is characterized above, then the minimizing coefficients are precisely those in section 3 since, if $X_1, X_2, \ldots, X_n$ are independent with lower bound φ and ν characterizes the lower tail of f, then $-X_1, -X_2, \ldots, -X_n$ are independent with upper bound $-\varphi$ and the upper tail of the distribution of $-X_i$ is characterized by ν. Finally, the largest r order statistics based on $-X_1, -X_2, \ldots, -X_n$ are the negatives of the smallest r order statistics based on $X_1, X_2, \ldots, X_n$.

References

Cooke, P.J. (1979). Statistical inference for bounds of random variables. Biometrika. To appear.

Gnedenko, B. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. Ann. Math., 44, 423-454.

Robson, D.S. and Whitlock, J.H. (1964). Estimation of a truncation point. Biometrika, 51, 33-39.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
37

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (*and Subtitle*)
Optimal Linear Estimation of Bounds of Random Variables

5. TYPE OF REPORT & PERIOD COVERED
TECHNICAL REPORT

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(*s*)
Peter Cooke

8. CONTRACT OR GRANT NUMBER(*s*)
DAAG29-77-G-0031

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Statistics
Stanford University
Stanford, CA 94305

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
P-14435-M

11. CONTROLLING OFFICE NAME AND ADDRESS
U. S. Army Research Office
Post Office Box 12211
Research Triangle Park, NC 27709

12. REPORT DATE
September 24, 1979

13. NUMBER OF PAGES
8

14. MONITORING AGENCY NAME & ADDRESS(*if different from Controlling Office*)

15. SECURITY CLASS. (*of this report*)
UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (*of this Report*)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (*of the abstract entered in Block 20, if different from Report*)

18. SUPPLEMENTARY NOTES
The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents. This report partially supported under Office of Naval Research Contract N00014-76-C-0475 (NR-042-267) and issued as Technical Report No. 276.

19. KEY WORDS (*Continue on reverse side if necessary and identify by block number*)

Linear estimator; Gnedenko's condition; Mean squared error; Asymptotic relative efficiency.

20. ABSTRACT (*Continue on reverse side if necessary and identify by block number*)

PLEASE SEE REVERSE SIDE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (*When Data Entered*)

OPTIMAL LINEAR ESTIMATION OF BOUNDS OF RANDOM VARIABLES

The problem of estimating the bounds of random variables has been discussed in Cooke (1979). Here we discuss optimality of estimates when the data is censored so that only the r largest or smallest of the observations is available for estimating a bound. For fixed r we find the linear function of the censored data which is the optimal estimator of a bound in the sense that, when the sample size is large, the estimator has smallest mean squared error among all such linear estimators. Provided r is not close to one, these estimators are almost optimal when the entire sample is available since, for example, when estimating an upper bound and the sample size is large, the largest few observations carry most of the information about the bound. This fact is illustrated in one case.

276/37